Analysis on 2022 Major League Baseball Data

Indo Badial, RJ Burson, and

Melissa Gomez

California State University, Long Beach

STAT 550 - Multivariate Statistical Analysis

15 December 2022

**Introduction**

The purpose of our project is to analyze 2022 Major League Baseball hitting statistics in order to predict a players on base plus slugging percentage over the course of a 162 game season. In order to do so, we use techniques such as principal component analysis to reduce the dimension of our dataset, inferences on the mean vector, factor analysis to explain any covariance/correlation structures amongst variables, and cluster analysis to possibly reduce the number of observations by grouping similar individuals together. Our data was collected from Baseball Savant, a site dedicated to providing Statcast metrics, and advanced statistics. Our dataset includes metrics such as [On-base Plus Slugging (OPS),](https://www.mlb.com/glossary/standard-stats/on-base-plus-slugging#:~:text=Definition,referred%20to%20as%20OPS%20against.) [Exit Velocity (EV)](http://m.mlb.com/glossary/statcast/exit-velocity), and [Launch Angle (LA)](http://m.mlb.com/glossary/statcast/launch-angle). The dataset consists of 469 observations (players) and 25 observable variables.

**Methods and Analysis**

The first step in the process was to perform exploratory data analysis to learn more about the nature of the dataset. To accomplish this, a correlation matrix between all 25 variables was produced and analyzed to see if any variables were highly correlated with each other. This would mean that several variables essentially give us the same information about OPS, meaning data reduction techniques should be used for simplicity.

Table

Description automatically generated

The simple statistics of the original full model are given in the table to the right. The correlation matrix is very large (25x25 size matrix) so only the code for it is included in the appendix. However, there were several variables with high correlation (>.80) which justify the use of methods such as Principal Component Analysis and Factor Analysis used later.

Principal component analysis allows us to reduce our large datasets with many variables into a smaller dataset with few variables. Thus, for the purpose of data reduction, we use principal analysis on our dataset. Since O[n-base Plus Slugging (OPS)](https://www.mlb.com/glossary/standard-stats/on-base-plus-slugging#:~:text=Definition,referred%20to%20as%20OPS%20against.) is our dependent variable we exclude it from our analysis. The first step of principal component analysis is to analyze the eigenvalues of the correlation matrix; the larger eigenvalues are extracted first and will range from 1-24 because there are 24 other observable variables aside from OPS, as shown in Figure 1. In the figure, the 6 largest eigenvalues are 6.25, 4.00, 2.78, 2.31, 1.50, and 1.33 which together account for 75.72% of the standardized variance. Only the first 6 components are used on the criterion of the eigenvalues-greater-than-one rule; in Figure 1 we see that the 7th eigenvalue is just below 1 at 0.97. A useful visual aid that helps determine an appropriate number of principal components to use is the scree plot. We look for an elbow/bend in the scree plot to help determine the appropriate number of principal components. In Figure 2, we see that there are a couple “elbows” that occur at about i=3 and i=5 but based on what we have discussed, 6 principal components provide a sufficient summary of our data.

Chart

Description automatically generated  
Table

Description automatically generated  
  
  
  
 **Figure 2.** Scree Plot and Variance Explained Plot

**Figure 1.** Eigenvalues of the Correlation Matrix

We extend our analysis to Factor Analysis, to help us determine the pattern of the relation among the variables. In Figure 3, we have the factor pattern also known as the factor loading matrix where each element in the matrix is called factor loadings. The factor loadings tell us the strength of the relation between the variables and the components. We can see that Factor1 and Factor2 have a significant correlation to many variables and together account for 42.74% of the variance. Next, this reduced model, that preserves much of the data’s variability, is used for cluster analysis and grouping certain hitters together based on similar characteristics.

**Table

Description automatically generated**

**Figure 3.** Factor Loading Matrix

Chart

Description automatically generated with low confidence

**Figure 4.** Plot of Factor1 and Factor2

After analyzing the factor pattern of the factor analysis, a couple of the factors are very similar to each other and are not needed. Factor 1 and Factor 6 are very similar to each other because a player with a high value in each of these factors is a player who hits the ball in the air a lot. A player with a low value in Factor 1 or 6 is a player who hits the ball on the ground a lot. Factor 2 and Factor 3 were also similar to each other. Players with a high value in Factors 2 and 3 were patient hitters who made a lot of good contact. Players with a low value in these factors were hitters who swung the bat a lot and didn’t make very good contact when doing so.

The cluster analysis moving forward was done just using Factors 1, 2, 4 and 5. If a player had a high value in Factor 4 it meant that the batter was thrown a lot of pitches out of the strike zone and swung at a lot of pitches out of the strike zone. If a batter had a low Factor 4 value, it is because they saw more pitches in the zone and didn’t swing at the pitches out of the zone.  Factor 5 is contrasting how the batters spray the ball around the field. A high value in Factor 5 indicates that the batter pulls the ball more often whereas a player with a low value is someone who hits the ball to all fields.

When using proc cluster in SAS, Ward's minimum variance method was applied and the tree plot in Figure 5 helped determine how many clusters should be used.

Chart

Description automatically generated

**Figure 5.** Tree Plot

From this tree plot it was determined to use a number of 3 clusters because the more clusters that were added the less interpretable the clusters became. They all became intermixed with one another, and it was hard to differentiate one cluster from another. The next step was to plot these clusters onto a plot with Factor1 on the x-axis and Factor2 on the y-axis that can be seen in Figure 6. Another plot in Figure 7 shows the clusters on a plot with Factor1 on the x-axis and Factor4 on the y-axis.

Chart, scatter chart

Description automatically generated

**Figure 6.** Factor1\*Factor2

In Figure 6 all three of the clusters seem to have hitters spread across the Factor 1 values. Meaning there are a variety of different types of hitters in each cluster(heavy ground ball hitter, heavy fly ball hitter, and in between). The clusters looked grouped by tiers when it comes to Factor 2. Which means they were grouped by how patient they were at the plate.  Cluster 3(red) is grouped toward the bottom of Factor 2 values being the least patient, cluster 1(black) is grouped near the middle values of Factor 2, and cluster 2(blue) is towards the top of the graph with the highest Factor 2 values being the most patient. This is backed up by the fact that Joey Gallo at the top of the graph saw 4.2 pitches per plate appearance in 2022 and Nick Madrigal at the bottom of the graph saw 3.81 pitches per plate appearance in 2022. The league average was around 3.95.

**Conclusion**

As shown in the analysis, the issue of creating a sufficient prediction model for the on-base plus slugging of MLB batters requires extensive data reduction as a result of high correlation amongst variables. This was done through the use of principal component analysis which retained 75.72% of the standardized variance of the data. In addition to this, the factor analysis yielded a final model that consisted of only 6 factors as opposed to 25 original variables. This factor model was then used for more efficient cluster analysis used to group similar players together. As seen in Figure 6, there are different types of successful hitters and different types of bad hitters in baseball. So, it isn't easy to group by a hitter's tendency to hit the ball in the air or on the ground because there are good and bad hitters with each tendency. It is easier to group by how patient the hitters are at the plate, however.

**References**

“Factor Analysis.” *Statistics Solutions*, 10 Aug. 2021, https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/factor-analysis/.

Johnson, Richard Arnold, and Dean W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2007.

**Code**

FILENAME REFFILE "C:/Users/014497819/Downloads/stats (7).csv";

PROC IMPORT DATAFILE=REFFILE

DBMS=CSV

OUT=BASEBALL;

GETNAMES=YES;

RUN;

DATA BASEBALL; \*renaming vars for simplicity;

RENAME exit\_velocity\_avg=AvergaeExitVelocity

launch\_angle\_avg=AverageLaunchAngle

barrel\_batted\_rate=BarrelBattedRate

solidcontact\_percent=SolidContactPct

flareburner\_percent=FlareBurnerPct

poorlyunder\_percent=PoorlyUnderPct

poorlytopped\_percent=PoorlyToppedPct

poorlyweak\_percent=PoorlyWeakPct

hard\_hit\_percent=HardHitPct

z\_swing\_percent=ZSwingPct

oz\_swing\_percent=OZSwingPct

oz\_contact\_percent=OZContactPct

out\_zone\_percent=OutZonePct

iz\_contact\_percent=IZContactPct

in\_zone\_percent=InZonePct

edge\_percent=EdgePct

whiff\_percent=WhiffPct

pull\_percent=PullPct

straightaway\_percent=StraightAwayPct

opposite\_percent=OppositePct

groundballs\_percent=GroundballsPct

flyballs\_percent=FlyballsPct

linedrives\_percent=LinedrivesPct

popups\_percent=PopupsPct;

SET BASEBALL;

full\_name=cat(' ',\_first\_name , last\_name);

RUN;

\*Cleaning up data set;

DATA BASEBALL;

set BASEBALL (drop = last\_name \_first\_name player\_id  year VAR30);

RUN;

/\*\*\*\*\*\*\*\* Correlation Matrix \*\*\*\*\*\*\*/

proc corr data=BASEBALL;

run;

\*Factor Analysis to determine how many factors to use;

PROC FACTOR DATA=BASEBALL SIMPLE CORR;

VAR AvergaeExitVelocity

AverageLaunchAngle BarrelBattedRate SolidContactPct FlareBurnerPct PoorlyUnderPct

PoorlyToppedPct PoorlyWeakPct HardHitPct ZSwingPct OZSwingPct OZContactPct OutZonePct

IZContactPct InZonePct EdgePct WhiffPct PullPct StraightAwayPct OppositePct GroundballsPct

FlyballsPct LinedrivesPct PopupsPct;

RUN;

PROC PRINCOMP DATA=BASEBALL OUT=PRINCOMP;

VAR AvergaeExitVelocity

AverageLaunchAngle BarrelBattedRate SolidContactPct FlareBurnerPct PoorlyUnderPct

PoorlyToppedPct PoorlyWeakPct HardHitPct ZSwingPct OZSwingPct OZContactPct OutZonePct

IZContactPct InZonePct EdgePct WhiffPct PullPct StraightAwayPct OppositePct GroundballsPct

FlyballsPct LinedrivesPct PopupsPct;

RUN;

\*factor analysis without OPS var;

\*Determined to use 6 factors;

Proc Factor DATA=BASEBALL METHOD=prin NFACT=6 out=baseballfact

     ROTATE=NONE PREPLOT PLOT; \*NO ROTATION;

var AvergaeExitVelocity

AverageLaunchAngle BarrelBattedRate SolidContactPct FlareBurnerPct PoorlyUnderPct

PoorlyToppedPct PoorlyWeakPct HardHitPct ZSwingPct OZSwingPct OZContactPct OutZonePct

IZContactPct InZonePct EdgePct WhiffPct PullPct StraightAwayPct OppositePct GroundballsPct

FlyballsPct LinedrivesPct PopupsPct;

run;

/\*\*\*\* Decided to use only Factors 1,2,4&5 for clustering because after interpreting the factors Factor3 was similar to

factor 2 and factor 6 was similar to factor 1\*/

Proc Cluster Data=baseballfact Method=ward OutTree=baseballtree pseudo;

  Var Factor1 Factor2 Factor4 Factor5; \*all of the variables to be used in cluster analysis;

 Id full\_name;

Run;

\* Method options: single, complete, average, centroid, ward, etc. ;

GOptions Reset=Symbol Reset=Axis;

Proc GPlot Data=baseballtree;

 Plot \_HEIGHT\_\*\_NCL\_=1 / VAxis=Axis1;

 Axis1 Label=(A=90);

 Symbol1 C=Black V=Dot I=SplineS;

Run;

Quit;

Proc Gplot;

 plot \_PSF\_\*\_NCL\_=1;

Axis1 Label=(A=90);

 Symbol1 C=Black V=Dot I=SplineS H=.34;

run;

Proc Tree Data=baseballtree NCL=3 out=clusters VAxis=Axis1;

 Id full\_name;

 Axis1 Label=(A=90);

Run;

\*Merge the datasets with clusters and the dataset with factors;

Proc Sort data=baseballfact;

By full\_name; run;

Proc sort data=clusters;

by full\_name; run;

data baseballclust;

merge baseballfact clusters;

by full\_name;

drop ClusName; run;

/\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* Plot FACTOR1 and FACTOR2 \*\*\*\*\*\*\*\*\*/

Proc Gplot data=baseballclust;

plot Factor1\*Factor2=cluster / VAxis=Axis1 HAxis=Axis2;

 Axis1 Label=(A=90 "Factor 2")

       Order=(-4 To 4 By 1)

       Length=5.75in;

 Axis2 Label=("Factor 1")

       Order=(-4 To 4 By 1)

       Length=5.75in;

 Symbol1 C=Black V=Circle    I=None Pointlabel=(C=Black H=0.75 "#full\_name");

 Symbol2 C=Blue  V=Triangle  I=None Pointlabel=(C=Blue  H=0.75 "#full\_name");

 Symbol3 C=Red   V=Square    I=None Pointlabel=(C=Red   H=0.75 "#full\_name");

Run;

Quit;